NeuroImage 94 (2014) 287–302

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Fast and accurate modelling of longitudinal and repeated measures neuroimaging data

Bryan Guillaume [a,b,c], Xue Hua [d], Paul M. Thompson [d], Lourens Waldorp [e], Thomas E. Nichols [b,f,g,*], for the Alzheimer's Disease Neuroimaging Initiative [1]

[a] Cyclotron Research Centre, University of Liège, 4000 Liège, Belgium
[b] Department of Statistics, University of Warwick, Coventry, UK
[c] Global Imaging Unit, GlaxoSmithKline, Stevenage, UK
[d] Imaging Genetics Center, Laboratory of Neuro Imaging, Dept. of Neurology & Psychiatry, UCLA School of Medicine, Los Angeles, CA 90095, USA
[e] Department of Psychological Methods, University of Amsterdam, Amsterdam, The Netherlands
[f] Warwick Manufacturing Group, University of Warwick, Coventry, UK
[g] Oxford Centre for Functional MRI of the Brain, University of Oxford, Oxford, UK

ARTICLE INFO

Article history:
Accepted 10 March 2014
Available online 18 March 2014

Keywords:
Longitudinal Modelling
Sandwich Estimator
Marginal Modelling
ADNI

ABSTRACT

Despite the growing importance of longitudinal data in neuroimaging, the standard analysis methods make restrictive or unrealistic assumptions (e.g., assumption of Compound Symmetry—the state of all equal variances and equal correlations—or spatially homogeneous longitudinal correlations). While some new methods have been proposed to more accurately account for such data, these methods are based on iterative algorithms that are slow and failure-prone. In this article, we propose the use of the Sandwich Estimator method which first estimates the parameters of interest with a simple Ordinary Least Square model and second estimates variances/covariances with the "so-called" Sandwich Estimator (SwE) which accounts for the within-subject correlation existing in longitudinal data. Here, we introduce the SwE method in its classic form, and we review and propose several adjustments to improve its behaviour, specifically in small samples. We use intensive Monte Carlo simulations to compare all considered adjustments and isolate the best combination for neuroimaging data. We also compare the SwE method to other popular methods and demonstrate its strengths and weaknesses. Finally, we analyse a highly unbalanced longitudinal dataset from the Alzheimer's Disease Neuroimaging Initiative and demonstrate the flexibility of the SwE method to fit within- and between-subject effects in a single model. Software implementing this SwE method has been made freely available at http://warwick.ac.uk/tenichols/SwE.

Introduction

Longitudinal data analysis is of increasing importance in neuroimaging, particularly in structural and functional MRI studies. There were over 1000 publications in 2012 to mention "longitudinal fMRI", which is 3.9% of all "fMRI" 2012 publications and up from 1.5% in 2000.[2] Unfortunately, while the current versions of the two most widely used packages (i.e. SPM and FSL) are computationally efficient, when they model more than two time points per subject they must make quite restrictive assumptions. In particular, FSL v5.0 must assume Compound Symmetry, a simple covariance structure where the variances and correlations of the repeated measures are constant over time, and a fully balanced design. SPM12 unrealistically assumes a common longitudinal covariance structure for the whole brain. This motivates recent publications proposing methods to better model neuroimaging longitudinal data (Bernal-Rusiel et al., 2013a, 2013b; Chen et al., 2013; Li et al., 2013; Skup et al., 2012), however, all of these methods entail iterative optimisation at each voxel.

In neuroimaging, the two most widely longitudinal approaches currently used are the Naïve Ordinary Least Squares (N-OLS) modelling and the Summary Statistics Ordinary Least Squares (SS-OLS) modelling. The N-OLS method tries to account for the intra-visit correlations existing in the data by including subject indicator variables (i.e. an intercept per

* Corresponding author at: Department of Statistics, University of Warwick, Coventry, CV4 7AL, United Kingdom.

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

[2] Based on Pubmed searches of "longitudinal AND fMRI" in all fields, versus just "fMRI". This is a crude measure, but does reflect the growing role of this type of study.

http://dx.doi.org/10.1016/j.neuroimage.2014.03.029

subject) in an OLS model. This approach is fast, but does not allow one to make valid inferences on pure between-subject covariates (e.g., group intercept or gender) and is valid only under a balanced design and Compound Symmetry (CS). The SS-OLS method proceeds by first extracting a summary statistic of interest for each subject (e.g., slope with time) and then uses a group OLS model to infer on the summary measures. This method is also fast and has the advantage of reducing the analysis of correlated data to an analysis of independent data, but this summary data may be highly variable as it is based on single-subject fits. In the context of one-sample t-tests, Mumford and Nichols (2009) showed that this approach is robust under heterogeneity, but warned that it is probably not the case for more general regression models.

In biostatistics, the analysis of longitudinal data is a long-standing problem and is generally performed by using either Linear Mixed Effects (LME) models or marginal models. The LME models include random effects which account for the intra-visit correlations existing in the data. Nevertheless, they require iterative algorithms which are generally slow and may fail to converge to a correct solution. Another issue with LME models is the complexity of specifying and fitting the model. For example, the random effects and the covariance structure of the error terms need to be specified (e.g., only random intercepts? Also random slopes?) and, unfortunately, a misspecification of those may lead to invalid results. These are particularly serious problems in neuroimaging as model assessment is difficult and a single model must be used for the whole brain. As a consequence, the use of LME models in neuroimaging may be prohibitively slow, and may lead to statistical images with missing or invalid results for some voxels in the brain. To limit the convergence issues, one may be tempted to use a LME model with only a random intercept per subject. Unfortunately, like the N-OLS model, this model assumes CS which is probably not realistic, especially for long studies carried out over years and with many visits. In contrast, the marginal modelling approach implicitly accounts for random effects, treats the intra-visit correlations as a nuisance and focuses the modelling only on population averages. They have appealing asymptotic properties, are robust against model misspecification and, as there are no explicit random effects, are easier to specify than LME models. However, they only focus on population-averaged inferences or predictions, typically require iterative algorithms and assume large samples.

Recently, Bernal-Rusiel et al. (2013a) proposed the use of LME models to analyse longitudinal neuroimaging data, but only on a small number of regions of interest or biomarkers, Chen et al. (2013) and Bernal-Rusiel et al. (2013b) extended the use of the LME models to mass-univariate settings. In particular, Bernal-Rusiel et al. (2013b) proposed the use of a spatiotemporal LME method based on a parcellation of the brain into homogeneous areas; in each area, they model the full spatiotemporal covariance structure by assuming a common temporal covariance structure across all the points and a simple spatial covariance structure. Skup et al. (2012) and Li et al. (2013) proposed to use marginal models to analyse neuroimaging longitudinal data. Specifically, Skup et al. (2012) proposed a Multiscale Adaptive Generalised Method of Moments (MA-GMM) approach which combines a spatial regularisation method with a marginal model called Generalised Methods of Moments (GMM; Hansen, 1982; Lai and Small, 2007) and Li et al. (2013) proposed a Multiscale Adaptive Generalised Estimating Equations (MA-GEE) approach which also combines a spatial regularisation method, but with a marginal model called Generalised Estimating Equations (GEE; Liang and Zeger, 1986). Thanks to their appealing theoretical asymptotic properties, the two latter methods seem very promising for analysing longitudinal neuroimaging data. Nevertheless, like the LME models, they require iterative algorithms, which make them slow, and – due to the fact that they rely on asymptotic theoretical results – their use may be problematic in small samples.

In this paper, we propose an alternative marginal approach. We use a simple OLS model for the marginal model (i.e. no subject indicator variables) to create estimates of the parameters of interest. For standard errors of these estimates, we use the so-called Sandwich Estimator (SwE; Eicker, 1963) to account for the repeated measures correlation. The main property of the SwE is that, under weak conditions, it is asymptotically robust against misspecification of the covariance model. In particular, this robustness allows us to combine the SwE with a simple OLS model which has no covariance model. Thus, this method is easy to specify and, with no need for iterative computations, is fast and has no convergence issues. Moreover, the proposed method can deal with unbalanced designs and heterogeneous variances across time and groups (or even subjects; more below on this). In addition, note that the SwE method can also be used for cross-sectional designs where repeated measures exist, such as fMRI studies where multiple contrasts of interests are jointly modelled, or even for family designs where subjects from the same family cannot be assumed independent. Nevertheless, like the MA-GMM and MA-GEE methods, the SwE method relies on asymptotic theoretical results, guaranteeing accurate inference only in large samples. Therefore, we also review and propose small sample adjustments that improve its behaviour in small samples.

The remainder of this paper is organised as follows. Starting from the LME model and its implied marginal model, we introduce the SwE method in its standard form. Then, we review and propose different adjustments to the SwE in order to improve its behaviour, mainly in the case of small samples. Finally, we assess the SwE method with intensive Monte Carlo simulations in a large range of settings and, more particularly, by analysing real brain images acquired as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI; Mueller et al., 2005).

Methods

The Linear Mixed Effects model and the marginal model

Using the formulation of Laird and Ware (1982), the LME model for individual i is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \tag{1}$$

where y_i is a vector of n_i observations for individual $i = 1,2,\ldots,m$, β is a vector of p fixed effects which is linked to y_i by the $n_i \times p$ design matrix X_i, b_i is a vector of r individual random effects which is linked to y_i by the $n_i \times r$ design matrix Z_i, and ϵ_i is a vector of n_i individual error terms which is normally distributed with mean 0 and covariance Σ_i. The individual random effects b_i are also normally distributed, independently of ϵ_i, with mean 0 and covariance D. Typically, the p fixed effects might include an intercept per group, a linear effect of time per group, a quadratic effect of time per-group or per-visit measures effects like, in the case of Alzheimer's Disease, the MMSE (Mini-Mental State Examination) score. The r random effects usually include a "random intercept" for each subject (modelled by a constant in Z_i) and may also include a "random slope" for each subject.

Instead of posing a model for each subject consisting of (common) fixed and (individual) random components, we can fit a model with only fixed components and let the random components induce structure on the random error. This is the so-called marginal model, which, for subject i, has the form

$$y_i = X_i\beta + \epsilon_i^* \tag{2}$$

where the individual marginal error terms ϵ_i^* have mean 0 and covariance V_i. Typically, the covariance is taken to be unstructured, but if data arise as per the LME model specified above, then $V_i = \Sigma_i + Z_i D Z_i$. We will denote by X the grand design matrix, the $n \times p$ stacked matrix of the m X_i's, where $n = \sum_i n_i$ is the total number of observations.

In LME models, the randomness of the data is modelled by both the random effects b_i and the error terms ϵ_i. The random effects b_i have an important impact on the variance modelling and have to be chosen carefully. This makes LME models quite difficult to specify in practice. In contrast, in the marginal model, all the randomness is treated as a

nuisance and is modelled by the marginal error terms ϵ_i^*. Therefore, the marginal models do not require the specification of random effects, making them easier to specify than LME models. Moreover, the marginal models are more flexible because they only require that the V_i be positive semi-definite. In the LME models, both Σ_i and D have to be positive semi-definite which is more restrictive (Molenberghs and Verbeke, 2011; Verbeke and Molenberghs, 2009; West et al., 2006). However, the marginal models are only focused on population-averaged inferences and predictions, and do not offer the possibility to make inferences on random effects or to predict subject-specific profiles like LME models do. Nevertheless, subject-specific inferences or predictions are not generally of interest in longitudinal neuroimaging studies and therefore, a marginal approach should be sufficient to analyse the data (for inferences on random effects parameters, see Lindquist et al., 2012).

In both models, the fixed effects parameters are estimated by

$$\hat{\beta} = \left(\sum_{i=1}^{m} X_i' W_i X_i\right)^{-1} \sum_{i=1}^{m} X_i' W_i y_i \tag{3}$$

where W_i is the so-called working covariance matrix of individual i (Diggle et al., 1994; Liang and Zeger, 1986). If $W_i = \mathrm{I}$ the identity matrix, it is the Ordinary Least Squares (OLS) estimate. If $W_i = V_i^{-1}$, it is the Generalized Least Squares (GLS) estimate, the Uniform Minimum Variance Unbiased Estimate.

The covariance matrix of the fixed parameter estimates $\mathrm{var}\left\{\hat{\beta}\right\}$ is estimated by

$$S = \underbrace{\left(\sum_{i=1}^{m} X_i' W_i X_i\right)^{-1}}_{Bread} \underbrace{\left(\sum_{i=1}^{m} X_i' W_i \hat{V}_i W_i X_i\right)}_{Meat} \underbrace{\left(\sum_{i=1}^{m} X_i' W_i X_i\right)^{-1}}_{Bread}, \tag{4}$$

where $\hat{V}_i$ is an estimate of the subject covariance V_i. The central part of this estimate can be conceptualised as a piece of meat between two slices of bread, giving rise to the name of Sandwich Estimator (SwE). If $m^{-1}\sum_{i=1}^{m} X_i' W_i \hat{V}_i W_i X_i$ consistently[3] estimates $m^{-1}\sum_{i=1}^{m} X_i' W_i V_i W_i X_i$, the SwE converges asymptotically to the true covariance matrix $\mathrm{var}\left\{\hat{\beta}\right\}$, even if W_i is misspecified (Diggle et al., 1994; Eicker, 1963, 1967; Huber, 1967; White, 1980). For GLS with $W_i = \hat{V}_i^{-1}$, the first two terms of S cancel and only the rightmost term remains. For OLS with $W_i = I$, we obtain the simplest version of the SwE which was first introduced by Eicker (1963, 1967). Note that, in practice, other choices for W_i are considered by assuming a non-identity structure for W_i and parametrising it with a vector of parameters, which then has to be estimated (Diggle et al., 1994; Liang and Zeger, 1986). These alternative choices are motivated by the fact that, even if the use of $W_i = \mathrm{I}$ yields consistent estimates and has been shown to be almost as efficient as the GLS estimator in some settings (Liang and Zeger, 1986; McDonald, 1993), it may lead to a non-negligible loss of efficiency[4] that can be ameliorated by more complicated forms of W_i (Fitzmaurice, 1995; Zhao et al., 1992). In particular, Fitzmaurice (1995) shows that, in the context of clustered binary data, an important loss of efficiency may arise for within-cluster covariates when the within-cluster correlation is high. Nevertheless, Pepe and Anderson (1994) showed that using a non-diagonal working covariance matrix may lead to inaccurate estimates of $\hat{\beta}$ and, further, using a non-identity covariance matrix requires generally the use of iterative algorithms to estimate the covariance parameters. Finally, as shown in the subsection Construction of the design matrix below, the loss of efficiency can be limited by an appropriate construction of the design matrix. For all these reasons, in this paper, we only focus on the use of the identity for W_i. See, however, Li et al. (2013) for the use of non-diagonal working covariance matrix within the framework of neuroimaging data, and Pepe and Anderson (1994) on the validity of using such working covariance matrices.

In LME models, the elements of V_i are generally defined as functions of a set of covariance parameters θ such that $V_i = V_i(\theta)$. These covariance parameters θ are estimated by either Maximum Likelihood (ML) or Restricted Maximum Likelihood (ReML) and are used to construct an estimate of V_i (Harville, 1977). In the SwE, V_i is usually estimated from the residuals $e_i = y_i - X_i\hat{\beta}$ by

$$\hat{V}_i = e_i e_i' \tag{5}$$

(Diggle et al., 1994). In the literature, the corresponding SwE is often referred to as HC0 (see, e.g., Long and Ervin, 2000) where "HC" stands for "Heteroscedasticity Consistent" and "0" stands for the fact that no small sample adjustment (see subsection Small sample adjustments) is made. Following this numbering, in this paper, we will refer to the corresponding SwE as S_0.

To perform inference on a linear combination of the parameters, $\mathcal{H}_0 : C\beta = 0$, a Wald test is generally used:

$$T = \left(C\hat{\beta}\right)' \left(CSC'\right)^{-1} \left(C\hat{\beta}\right)/q \tag{6}$$

where C is a matrix (or a vector) defining the combination of the parameters (contrast) tested and q is the rank of C. In large samples, this Wald test follows a χ_q^2 distribution. In small samples, while the obvious choice is an F-distribution with q and $n-p$ degrees of freedom, we show in the subsection Small sample adjustments that this is not a good approximation of the true null distribution of T when the SwE method is used.

Construction of the design matrix

In longitudinal data, the covariates have generally a between-subject component and a within-subject component. In the ADNI study, for example, the *Age* covariate has a between-subject component which can be summarised by the subject mean $\overline{Age}_i$ and a within-subject component which can be summarised by the difference with the subject mean $Age - \overline{Age}_i$. Including only the *Age* covariate in the design matrix means that we implicitly assume that the effects on the response is the same for both components. Actually, the effects of each component can be very different and, as shown by Neuhaus and Kalbfleisch (1998), the assessment of the effect of such between/within-subject covariates on the response can be very misleading. Therefore, we follow the recommendation of Neuhaus and Kalbfleisch (1998) and systematically split this kind of covariates into between- and within-subject components and include both in the design matrix. Moreover, as shown in Table 1, this helps also to improve the efficiency of the SwE method when assuming an identity working covariance matrix. This result shows that splitting the *Age* covariate makes the SwE nearly as efficient as GLS. It also demonstrates the (well-known) importance of centring covariates when inference is made on the intercepts, as this can be of interest in longitudinal fMRI studies. As the only reason to use a nontrivial working covariance matrix is to improve efficiency, we find that these covariate-splitting results are a compelling reason to only consider an identity working covariance matrix, and hence, in this paper, we exclusively use $W_i = \mathrm{I}$.

Homogeneous SwE

The standard SwE estimates a separate V_i for each subject, based only on the residuals of the i-th subject (Eq. (5)). Nevertheless, if the studied population can be subdivided into n_G groups within which the subjects are sharing similar properties, we may assume that the variances and covariances over subjects within each group are actually

[3] An estimator of a parameter is said to be consistent if it converges in probability to the true value of the parameter. Here, this is the case if $lim_{m\to\infty} m^{-1} \sum_{i=1}^{m} X_i' W_i \left(\hat{V}_i - V_i\right) W_i X_i = 0$.

[4] The efficiency of a scalar estimator is the inverse of estimator variance.

Table 1
Impact of splitting covariates into separate within- and between-subject covariates. Ages for full 817 subjects ADNI dataset were used to construct 4 models: (1) *Intercept* and *Age*, (2) *Intercept* and centred *Age*, (3) *Intercept*, mean age per subject $\overline{Age}_i$, and intra-subject-centred age $Age-\overline{Age}_i$, and (4) *Intercept*, centred mean age per subject $\overline{Age}_i-\overline{Age}$, and intra-subject-centred age $Age-\overline{Age}_i$. The relative efficiency is shown for each model for 3 possible values of ρ, the common intra-visit correlation. Here, we define relative efficiency as the ratio between the variance of the GLS estimate and the variance of the SwE estimate.

		Relative efficiency		
Model	Covariate	$\rho = 0$	$\rho = 0.5$	$\rho = 0.95$
1	*Intercept*	1	0.88	0.40
	Age	1	0.88	0.40
2	*Intercept*	1	0.94	0.89
	$Age-\overline{Age}$	1	0.88	0.40
3	*Intercept*	1	0.92	0.87
	$\overline{Age}_i$	1	0.92	0.87
	$Age-\overline{Age}_i$	1	1.00	1.00
4	*Intercept*	1	0.94	0.89
	$\overline{Age}_i-\overline{Age}$	1	0.92	0.87
	$Age-\overline{Age}_i$	1	1.00	1.00

homogeneous (Pan, 2001). For instance, in the ADNI study, the whole population can be divided into 3 groups: the Normal control (N), Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) groups in which the subjects may be assumed to share the same variances and covariances. We argue that this is a reasonable assumption as virtually all standard longitudinal neuroimaging analyses assumes homogeneous variance over all subjects. Therefore, in this paper, we propose an alternative version of the SwE which relies on the assumption of a common covariance matrix V_{0g} for all the individuals belonging to group $g = 1,\ldots, n_G$. To estimate V_{0g}, the observations have to be firstly classified into k_g visit categories (homogeneous groups) consistently defined between subjects in group g. For example, in the ADNI study, the MCI subjects were scanned at 0, 6, 12, 18, 24 and 36 months allowing us to divide the observations into $k_{MCI} = 6$ visit categories. Then, defining $m_{gkk'}$ as the number of subjects in group g who have data at both visit k and k', e_{ik} as the residual of subject i at visit k and $\mathcal{I}(g, k, k')$ as the subset of subjects in group g who have data at both visit k and k', the k^{th} diagonal element of V_{0g} can then be estimated by

$$\left(\hat{V}_{0g}\right)_{kk} = \frac{1}{m_{gkk}} \sum_{i \in \mathcal{I}(g,k,k)} e_{ik}^2. \tag{7}$$

The off-diagonal elements of V_{0g} corresponding to the visits k and k' can be estimated by

$$\left(\hat{V}_{0g}\right)_{kk'} = \hat{\rho}_{0gkk'} \sqrt{\left(\hat{V}_{0g}\right)_{kk} \left(\hat{V}_{0g}\right)_{k'k'}} \tag{8}$$

where $\hat{\rho}_{0gkk'}$ is an estimate of the correlation at visits k and k' in the group g and which can be computed by

$$\hat{\rho}_{0gkk'} = \frac{\sum_{i \in \mathcal{I}(g,k,k')} e_{ik} e_{ik'}}{\sqrt{\left(\sum_{i \in \mathcal{I}(g,k,k')} e_{ik}^2\right)\left(\sum_{i \in \mathcal{I}(g,k,k')} e_{ik'}^2\right)}}. \tag{9}$$

Note that, due to the possible presence of missing data, $\hat{V}_{0g}$ may not be positive semi-definite and, as a consequence, may lead to inaccurate results. Therefore, in presence of missing data, we make a spectral decomposition of $\hat{V}_{0g}$ and check whether all the eigenvalues of $\hat{V}_{0g}$ are positive. If this is not the case, we set all the negative eigenvalues to zero and reconstruct $\hat{V}_{0g}$ with the new eigenvalues, ensuring that $\hat{V}_{0g}$ is positive semi-definite. Note also that we normalise with $1/m_{gkk}$ instead of the usual bias corrective term $1/(m_{gkk} - 1)$ as we consider this sort of bias correction with other small sample adjustments in the next subsection, Small sample adjustments. Thus, in this SwE version, each $\hat{V}_i$ corresponds to a subset of the corresponding common covariance matrix $\hat{V}_{0g}$ depending on the visits measured for subject i. If the assumption of a common covariance matrix over subjects in a same group is valid, then the V_i should be more efficiently estimated in comparison to the standard approach. Note that this new SwE version depends on the way the population is subdivided and has two extreme cases, one assuming a single group and the other considering m homogeneous groups, equivalent to the standard SwE. We differentiate the various SwE versions using subscripts and superscripts on S: superscripts refer to the use of groups, with S^{Hom} referring to the use of n_G homogeneous groups, and S^{Het} referring to the standard SwE where heterogeneous, per-subject covariance estimates are used; subscripts refer to different possible small sample adjustments, described in the next subsection.

Small sample adjustments

In small samples, it is well known that the use of the standard SwE S_0^{Het} (heterogeneous, standard SwE, no small sample adjustment) may lead to inaccurate inferences (Chesher and Jewitt, 1987; Long and Ervin, 2000; MacKinnon and White, 1985). There are two explanations for this effect. The first explanation is that, since S_0 uses the Maximum Likelihood Estimate for each V_i, it is generally biased and tends to make liberal inferences (i.e. inflated False Positive Rates). The second explanation is that, because the standard Wald test (Eq. (6)) does not account for the randomness in S, the sampling distribution of T has heavier tails than the usual χ_q^2 distribution. Therefore, a naïve use of S_0^{Het} with T following a χ_q^2 null distribution also gives liberal inferences. Those two issues have led several authors to propose different adjustments to improve the behaviour of the SwE in small samples.

The first improvements proposed in the literature were focused on the correction of the bias of the SwE. The simplest adjustments proposed consist of multiplying the raw residuals e_{ik} by a correction factor before using them to estimate V_i. There are three principal alternative estimates based on this approach: S_1^{Het} (Hinkley, 1977; MacKinnon and White, 1985), S_2^{Het} (Horn et al., 1975; MacKinnon and White, 1985) and S_3^{Het} (MacKinnon and White, 1985). Note that, in the SwE literature, they are often referred to as HC1, HC2 and HC3, respectively. S_1^{Het} consists of using the raw residuals e_{ik} multiplied by $\sqrt{n/(n-p)}$ instead of the raw residuals e_{ik}; S_2^{Het} consists of using the adjusted residuals $e_{ik}/(1 - h_{ik})^{1/2}$ (where h_{ik} is the diagonal element of the Hat matrix $X(X'X)^{-1}X'$ corresponding to the observation of subject i at visit k) instead of the raw residuals e_{ik}; and S_3^{Het} consists of using $e_{ik}/(1 - h_{ik})$ instead of the raw residuals e_{ik}. Here, we also propose to use these small sample adjustments to compute each $\hat{V}_{0g}$ in the homogeneous versions of the SwE.

Subsequently, other authors proposed another type of improvement, altering the null distribution of the Wald test to account for the additional variability of the SwE (Bell and McCaffrey, 2002; Fay and Graubard, 2001; Hardin, 2001; Kauermann and Carroll, 2001; Lipsitz et al., 1999; Mancl and DeRouen, 2001; Pan and Wall, 2002; Waldorp, 2009). Most of the proposed adjustments consist of using a t-distribution (or an F-distribution) instead of a Normal distribution (or a χ^2 distribution) for the statistical test null distribution. The challenge is then to correctly define the degrees of freedom of the distribution. Here, we propose to use an approximate test statistic and null distribution similar to the one proposed in Pan and Wall (2002):

$$\frac{\nu - q + 1}{\nu q} \left(C\hat{\beta}\right)' (CSC')^{-1} \left(C\hat{\beta}\right) \sim F(q, \nu - q + 1) \tag{10}$$

where ν is a degrees of freedom parameter that has to be estimated. The justification of the proposed test and details about the estimation of ν can be found in Appendix A. The proposed approximate test is valid

for both the standard Heterogeneous and modified homogeneous SwE versions, but generally yields different estimates for v. As explained in Appendix A, homogeneous versions of the SwE produce more precise estimates of v than heterogeneous versions, further motivating the use of the homogeneous SwE. Note that, for a contrast of rank $q = 1$, the test simply becomes

$$\frac{C\hat{\beta}}{\sqrt{CSC'}} \sim t(v). \tag{11}$$

Monte Carlo simulations

Intensive Monte Carlo simulations were used in R (R Core Team, 2013) to assess the SwE method and compare it to the N-OLS, LME and SS-OLS methods. A variety of realistic settings were considered (detailed below), with 10,000 realisations created for each setting.

Simulations I

As a first set of simulations, we considered a selection of balanced and unbalanced designs. We used balanced designs consisting of longitudinal data generated for sample sizes of $m = 12$, 25, 50, 100 or 200 subjects with 3, 5 or 8 visits for each subject (a total of $5 \times 3 = 15$ distinct sample sizes). The subjects were divided into two groups A and B of equal sizes (except for $m = 25$ where the group A and B had 13 and 12 subjects, respectively) and we considered models consisting of, for each group, an intercept, a linear effect of visit and a quadratic effect of visit using orthogonal polynomials. In addition to these 15 balanced designs, we also considered the unbalanced design corresponding to the real ADNI dataset described in subsection Real data analysis. In order to also assess the methods in an unbalanced design but with a smaller number of subjects, we also considered four subsets of the full ADNI dataset obtained by iteratively removing half of the subjects at random in each group, leading to smaller and smaller sample sizes ($m_N = 229$, 114, 57, 29 and 14; $m_{MCI} = 400$, 200, 100, 50 and 25; $m_{AD} = 188$, 94, 47, 24 and 12). For this real unbalanced data design, we considered models consisting of, for each group, an intercept, the centred mean age per subject $\overline{Age}_i - \overline{Age}$ (referred to as cross-sectional "age" effect), the intra-subject centred age $Age - \overline{Age}_i$ (referred to as longitudinal "visit" effect) and their interaction (referred to as "acceleration").

For each realised dataset, each observation was first generated independently from a standard Normal distribution $\mathcal{N}(0,1)$. Then, the data for each subject $y_i = \left(y_{i1}, \ldots, y_{ik}, \ldots, y_{in_i}\right)'$ was correlated according to one of four different types of intra-visit covariance structure by premultiplying y_i by a square-root factor of the desired covariance matrix. The four covariance structures were generated according to the two following equations:

$$\mathrm{var}(y_{ik}) = \alpha_g(1 + \gamma t_k) \tag{12}$$

$$\mathrm{corr}(y_{ik}, y_{ik'}) = \rho(1 - \psi|t_k - t_{k'}|), \tag{13}$$

where α_g allows for different variances in each group, γ allows the variance to vary with visit, t_k ($t_{k'}$, respectively) is the time of measurement at visit k (visit k'), ρ controls the constant correlation over time and $\psi > 0$ allows for a linear decrease of the correlation over time. Table 2 summarises the parameter values used for the four covariance structures in the simulations for both the balanced and unbalanced ADNI designs.

For null simulations, the data was used immediately after being correlated. For non-null simulations, a signal was added according to the (per-subject centred) effect of visit.

For a given realised dataset and a given design, each of the four estimation methods were used in turn. Using custom R functions, eight versions of the SwE were used: S_0^{Het}, S_1^{Het}, S_2^{Het}, S_3^{Het}, S_0^{Hom}, S_1^{Hom}, S_2^{Hom} and S_3^{Hom} where the homogeneous groups were defined as groups A and B for the balanced designs and Normal, MCI and AD groups for the real unbalanced designs (see subsections Homogeneous SwE and Small sample adjustments for descriptions about these SwE versions); the SwE design matrices included all the effects described at the beginning of this subsection; the Wald tests were performed according to Eq. (10) estimating v in two different ways: as proposed in Eq. (A.16) and also, naïvely, by $m - p_B$ where p_B is the number of pure between-subject covariates (having a constant value for each subject) included in the model (e.g., intercepts, cross-sectional age effect) leading to 16 different variants for the SwE approach. The N-OLS included per-subject dummy variables, and thus precluded the use of the age effect (as age is a linear combination of the dummy variables). The SS-OLS approach used per-subject models, with a design matrix extracted from the appropriate rows and columns of the SwE design matrices, and contrasts that extracted quantities equivalent to the contrasts of interest used with the other models; the final model used with the SS-OLS approach was always a one-measure-per-subject OLS model allowing to test group effects equivalent to the one tested with the other methods. For both the N-OLS and SS-OLS methods, the function `lm` of the `stats` R package was used to estimate the model parameters, their variances/covariances and the degrees of freedom used in the Wald tests (i.e. the number of observations minus the number of parameters present in the considered model). The functions lme from the R package `nlme` (Pinheiro et al., 2013) and `lmer` from the R package `lme4` (Bates et al., 2012) were used to fit the LME models with the SwE design matrices for the fixed effects and a random intercept per subject as random effect; note that, as suggested by one of the reviewers, richer LME models were assessed in a second set of simulations (see subsection Simulations II). As the `lme4` package did not propose any estimation for the degrees of freedom, we used the ones estimated by the `nlme` package (Pinheiro and Bates, 2000) for all the `nlme` and `lme4` Wald tests.

For each realisation and contrast, several Wald tests T were computed and compared to F-distributions at a nominal level of significance of 5%. For null dataset, each significant realisation was counted as a False Positive detection and was used to compute the expected False Positive Rates (FPRs) for each method. The FPR of a valid test does not exceed the nominal level, while an invalid or liberal test will have an FPR in excess of the nominal level. Using a Normal approximation to binomial counts over 10,000 realisations, an exact test (FPR = 5%) should have a FPR between (4.57%, 5.43%) with 95% probability. Non-null simulations allowed the estimation of power with the True Positive Rates (TPRs) for each method.

Table 2
Covariance parameter values used in the simulations; γ and Ψ are expressed as "per visit" for the balanced design and "per year" for the ADNI design.

		Covariance parameters							
Design	Covariance structure	α_A	α_B	α_N	α_{MCI}	α_{AD}	γ	ρ	Ψ
Balanced	CS	1	1	–	–	–	0	0.95	0
	Toeplitz	1	1	–	–	–	0	1	0.1
	Group heterogeneity	1	2	–	–	–	0	0	0
	Visit heterogeneity	1	1	–	–	–	1	0	0
ADNI	CS	–	–	1	1	1	0	0.95	0
	Toeplitz	–	–	1	1	1	0	1	0.2
	Group heterogeneity	–	–	1	2	3	0	0	0
	Visit heterogeneity	–	–	1	1	1	2	0	0

Simulations II

Following the suggestions of the reviewers, we performed an additional three sets of simulations. In this set simulations similar to the first set were used, but we also considered LME models with a random intercept and time effect (slope), and LME models with a random intercept, linear and quadratic time effects. For these simulations, the ADNI design and its subsets were considered with a residual error covariance structure consisting of a Toeplitz correlation and an increasing variance

over time obtained from the Eqs. (12) and (13) with parameters $\alpha_N = 1$, $\alpha_{MCI} = 1$, $\alpha_{AD} = 1$, $\gamma = 2$/year, $\rho = 1$ and $\psi = 0.2$/year. The SwE, N-OLS and SS-OLS methods were fitted as in the first set of simulations. The LME models were fitted using the `lme4` package in the same way as the first set of simulations, but, in addition, we used more advanced functions to determine the degrees of freedom for each Wald test. Specifically, we used the `vcovAdj` and `get_ddf_Lb` functions of the `pbkrtest` R package (Halekoh and Højsgaard, 2013) to compute the Kenward-Roger covariance matrix correction and the Kenward–Roger effective degrees of freedom (Kenward and Roger, 1997), respectively.

Simulations III

The third set of simulations focused on the power analysis of all the methods (so, also including the two richer LME models investigated in subsection Simulations II) under CS and Toeplitz covariance structures in the unbalanced ADNI design. The covariance structures were produced with the same parameters as in the first set of simulations (see Table 2).

Simulations IV

In this final set, we conducted an experiment recording the failure rates of the LME models. For this, we used the same settings as in the first set of simulations (see subsection Simulations I), but only recorded the number of times the functions `lme` and `lmer` did not converge to a solution. The LME models considered were the same as the ones investigated in the second set of simulations (see subsection Simulations II), but, in addition, we included a model with a random intercept and a Toeplitz covariance structure for the error terms. Note that the latter model was only fitted with the `nlme` package as the `lme4` package do not allow the specification of correlation structure for the error terms.

Box's test of Compound Symmetry

As mentioned in the Introduction section, CS is the key assumption that justifies the use of N-OLS or random-intercept LME models. To assess whether the assumption of CS holds, Box (1950) proposed a test based on the determinant of the covariance matrix. It does not, however, accommodate missing data. In the presence of missing data, we construct a CS test using the largest possible subset of the full dataset which has no missingness. Eqs. (7), (8) and (9) (assuming only one group) are used to produce, at each voxel, an estimate of the common covariance matrix which can then be tested through the Box's test of CS, producing an image of F-scores (or p-values). Next, this image can be thresholded using a multiple testing correction (e.g., False Discovery Rate) and, if any voxels survive the threshold, we can conclude that there is evidence of violation of the assumption of CS.

Real data analysis

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California – San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see www.adni-info.org.

The dataset analysed in this paper is a modified version of the dataset produced and detailed by Hua et al. (2013). In brief, the dataset in Hua et al. (2013) consisted on 3314 images obtained after applying Tensor Based Morphometry (TBM) on 3314 brain MRI scans from 229 healthy elderly Normal controls (age: 76.0 ± 5.0 years, 119 Male (M)/110 Female (F)), 400 individuals with amnestic MCI (age: 74.8 ± 7.4 years, 257 M/143 F), and 188 probable AD patients (age at screening: 75.4 ± 7.5 years, 99 M/89 F). As shown in Table 3, the subjects were scanned at screening and followed up at 6, 12, 18 (MCI only), 24, and 36 months (Normal and MCI only) with visits counts of 4.16 ± 1.21, 4.43 ± 1.61 and 3.14 ± 1.07 for the Normal, MCI and AD subjects, respectively. More precisely, 817 screening TBM images were produced by considering the 817 screening scans and a Minimal Deformation Target (MDT) image, obtained from the scans of 40 randomly selected Normal subjects, as baseline; 2497 longitudinal TBM images were produced by considering, for each subject, the follow-up scans and the corresponding screening scan as baseline. More details about this dataset can be found in Hua et al. (2013). The 2497 longitudinal TBM images measure change *relative* to each subject's screening and *not* absolute volume (relative to a template). Therefore, we modified them by multiplying them with their corresponding TBM screening image in order to produce 2497 TBM images reflecting the brain volumes relative to a common baseline, the MDT image. We considered these modified 2497 TBM images with the unchanged 817 screening TBM images as the dataset to be analysed.

The modified dataset was analysed by using the N-OLS, SS-OLS and SwE methods with the same design matrices as used in the simulations (see subsection Monte Carlo simulations). SPM8 was used for the N-OLS and SS-OLS methods and a homemade SPM8 plug-in was used for the SwE method.

Results

SwE versions comparison in very small samples

Here, and for all results, we summarise the immense volume of Monte Carlo simulations by selecting the subset of findings that conveys the typical behaviour exhibited by the methods. Exhaustive results can be found in the Web Supplementary Material. Fig. 1 shows typical results obtained for 12 variants of the SwE in very small sample settings for a balanced design (12 subjects) and the unbalanced ADNI design (51 subjects). The standard S^{Het} tends to be liberal with the use of a

Table 3
Numbers of subjects scanned at baseline (0 month) and follow-up (6, 12, 18, 24 and 36 months) for the Normal controls (N), Mild Cognitive Impairment (MCI) and Alzheimer's Disease (AD) subjects in the ADNI dataset.

Scanning time	N	MCI	AD	Total
0 month	229	400	188	817
6 months	208	346	159	713
12 months	196	326	138	660
18 months	-	286	-	286
24 months	172	244	105	521
36 months	147	170	-	317

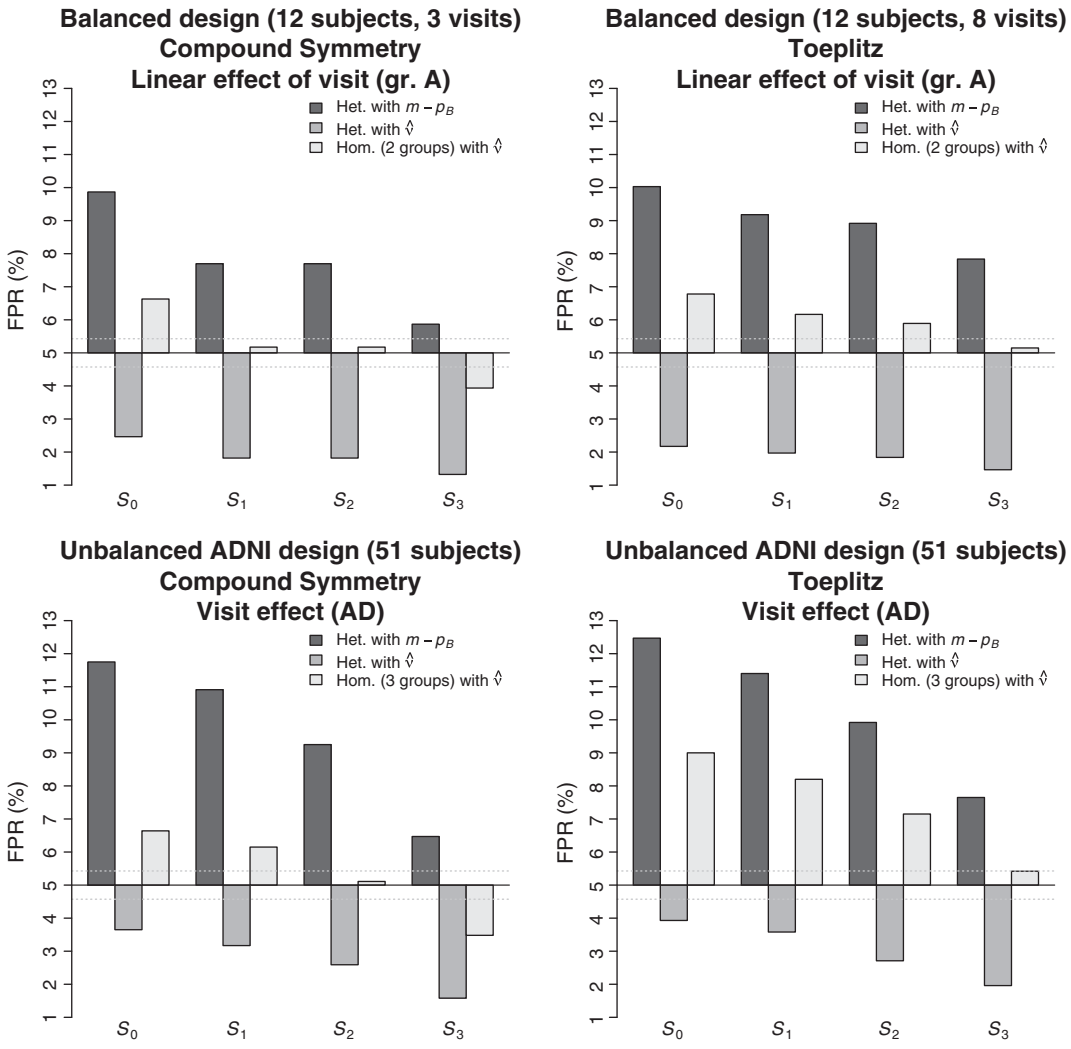

Fig. 1. FPR with different versions of the SwE in small samples with Compound Symmetry ($\rho = 0.95$) and Toeplitz ($\psi = 0.1$ per visit in the balanced design and $\psi = 0.2$ per year in the ADNI design) for the balanced and unbalanced ADNI design; all results are based on an F-test at nominal level 5%; S_0, S_1, S_2 and S_3 correspond to the SwE using the raw residuals e_{ik}, the adjusted residuals $e_{ik}\sqrt{N/(N-p)}$, the adjusted residuals $e_{ik}/(1-h_{ik})^{1/2}$ and the adjusted residuals $e_{ik}/(1-h_{ik})$, respectively; "Het. with $m - p_B$", "Het. with $\hat{v}$" and "Hom with $\hat{v}$" correspond to the standard heterogeneous SwE using (naïvely) $m - p_B$ as degrees of freedom, the standard heterogeneous SwE using the estimate proposed in Eq. (A.16) as degrees of freedom and the modified homogeneous SwE using the estimate proposed in Eq. (A.16) as degrees of freedom, respectively.

naïve estimation of v by $m - p_B$ (Fig. 1, dark grey bar) and conservative with the estimation of v by the estimate proposed in Eq. (A.16) (Fig. 1, medium grey bar). The homogeneous version (assuming homogeneity within groups) controls the FPR more accurately than the heterogeneous versions, with the better results obtained with the versions S_2^{Hom} and S_3^{Hom} (Fig. 1, light grey bar). Note that the settings selected in Fig. 1 were chosen in order to show some of the most severe adverse behaviour of these two versions, meaning that, they were, in general, controlling the FPR better in the simulations. Overall, S_2^{Hom} appears to be slightly liberal and S_3^{Hom} slightly conservative.

Methods comparison

FPR control

For the random-intercept LME models, we only show the results obtained with the `lme4` package as the results obtained with the `nlme` package were almost identical. Table 4 summarises qualitatively how the methods were able to control the FPR in the first set of simulations with different settings. The N-OLS method cannot provide inference on between-subject effects, but otherwise shows a performance similar to the random-intercept LME method. Specifically, the N-OLS and random-intercept LME methods struggle with variance heterogeneity (between groups or over time) and Toeplitz covariance structures, being either conservative or liberal depending on the setting. On between-subject effects, the random-intercept LME method has problems with variance heterogeneity. The SS-OLS method fares somewhat better than the N-OLS and random-intercept LME methods for balanced designs, but falls down on variance heterogeneity between groups and within-subject effects in the unbalanced design. Finally, with enough subjects, the SwE (S_3^{Hom}) seems accurate in all the settings, but, as shown in Figs. 1 and 2, it may slightly suffer from conservativeness in very small samples. Note that, as suggested by one of the reviewers, we also simulated a SwE assuming a common covariance matrix for all the subjects (one group in the modified homogeneous SwE) and found, under heterogeneous group variances, similar poor behaviours as the three other methods. See Web Supplementary Material for additional quantitative results comparing the methods.

Regarding the second set of simulations, the N-OLS, SS-OLS, random-intercept LME and SwE methods exhibited similar behaviours to the ones observed in the first set of simulations under variance heterogeneity over time. The LME models with a random intercept and a random effect of time per subject, and the LME models with a random intercept, a random effect of time and a random quadratic effect of time per subject had similar results to the ones of the SwE (S_3^{Hom}) method and

Table 4
Summary of simulation results for the False Positive Rate (FPR) control in different covariance settings, for between- and within-subject effects, and in the balanced and unbalanced ADNI designs. "R.-int." stands for Random-intercept; "n/a" stands for not applicable indicating that this type of inference is not possible for that particular method; "●" stands for an accurate FPR control, "+" for an invalid (liberal) FPR control, "−" for a conservative FPR control, "+/−" for both behaviours; "++/−" indicates an FPR control that is generally invalid, but also sometimes conservative; and "●/−" stands for an FPR control sometimes slightly conservative in small sample settings ($m < 50$ in the balanced design and $m < 200$ in the unbalanced ADNI design) but accurate otherwise. See Web Supplementary Material for detailed quantitative results.

Design	Cov. type	Effect type	N-OLS	R.-int. LME	SS-OLS	SwE (S_3^{Hom})
Balanced	CS	Between	n/a	●	●	●/−
		Within	●	●	●	●/−
	Toeplitz	Between	n/a	●	●	●/−
		Within	++/−	++/−	●	●/−
	Het. groups	Between	n/a	+/−	+/−	●/−
		Within	+/−	+/−	+/−	●/−
	Het. visits	Between	n/a	●/−	●	●/−
		Within	●	●	●	●/−
Unbalanced (ADNI)	CS	Between	n/a	●	●	●
		Within	●	●	+/−	●/−
	Toeplitz	Between	n/a	●	●	●
		Within	+	+	+/−	●
	Het. groups	Between	n/a	+/−	+/−	●
		Within	+/−	+/−	+/−	●
	Het. visits	Between	n/a	−	●	●
		Within	+	+	+/−	●

seemed accurate for all the settings. Note that in the third set of simulations, only the SwE (S_3^{Hom}) seemed to be able to control the FPR accurately under a Toeplitz covariance structure. In particular, as it can be seen in the Web Supplementary Material, the 2 richer LME models seemed liberal (e.g., in the full ADNI design and testing for a difference of visit effect between AD and MCI subjects at 5% level of significance, the LME model with a random intercept and a random effect of time per subject had a FPR of 6.1% while the LME model with a random intercept, a random effect of time and a random quadratic effect of time per subject had a FPR of 7.4 %).

Power analysis

Power comparisons are only interpretable when the methods considered control the FPR. Thus, as the majority of the compared methods had issues to control the FPR under Toeplitz covariance or variance heterogeneity (between groups, over time), we only show power comparisons for CS. Note that a comparison in the Toeplitz case can be found in the Web Supplementary Material.

Fig. 3 shows the results of the power analysis for a greater visit effect in AD relative to MCI subjects under the assumption of CS obtained from the third set of simulations (see subsection Simulations III). The SwE method is less powerful than the N-OLS and LME methods with a difference of power larger in very small samples, but becoming narrower and narrower when the sample size increases. Finally, even if the SS-OLS method is liberal for the FPR control (see Fig. 3, top left), it seems clearly

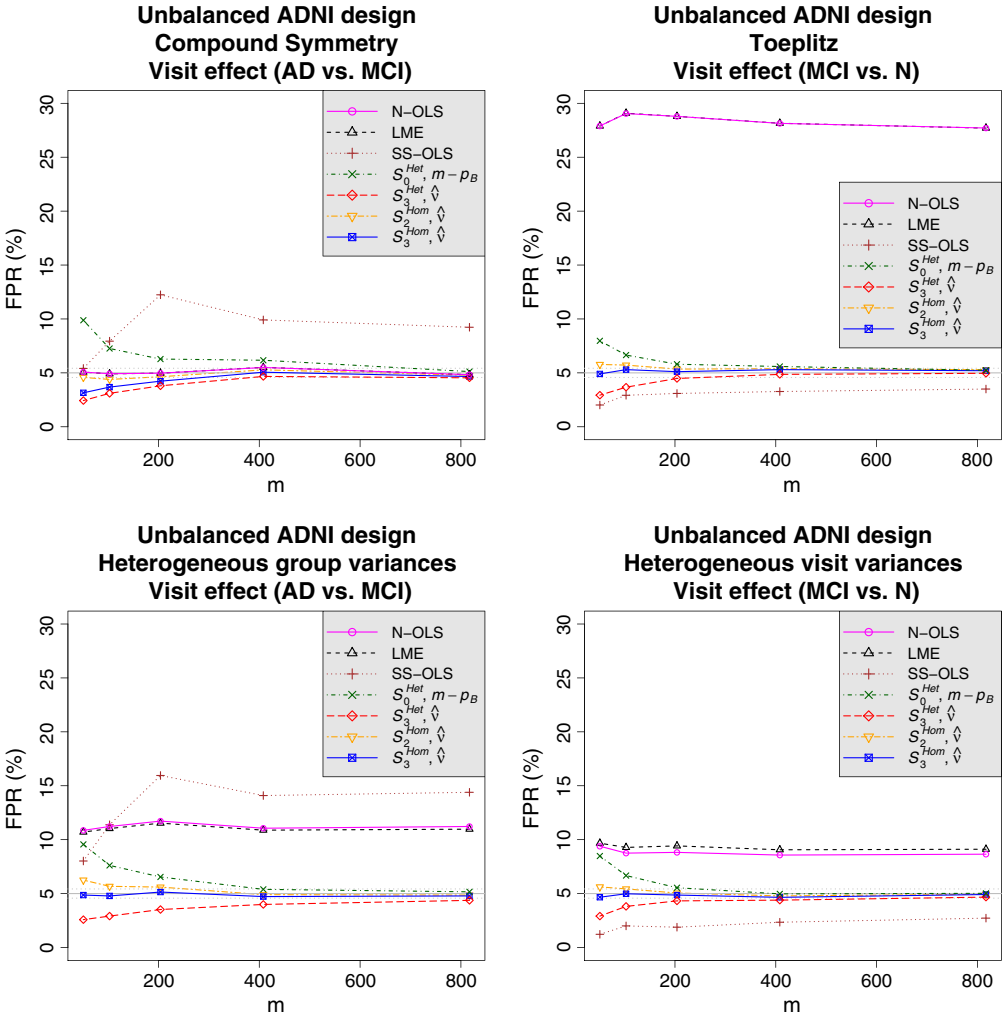

Fig. 2. FPR comparison on the visit effect with Compound Symmetry (top left, $\rho = 0.95$), Toeplitz (top right, $\psi = 0.2$ per year), heterogeneous group variances (bottom left, $\alpha_N = 1$, $\alpha_{MCI} = 2$ and $\alpha_{AD} = 3$) and heterogeneous visit variances (bottom right, $\gamma = 2$ per year) for the ADNI design. All results are based on an F-test at nominal level 5%, and the results for the LME method correspond to the random-intercept LME model. See Fig. 1 for a description of the SwE versions.

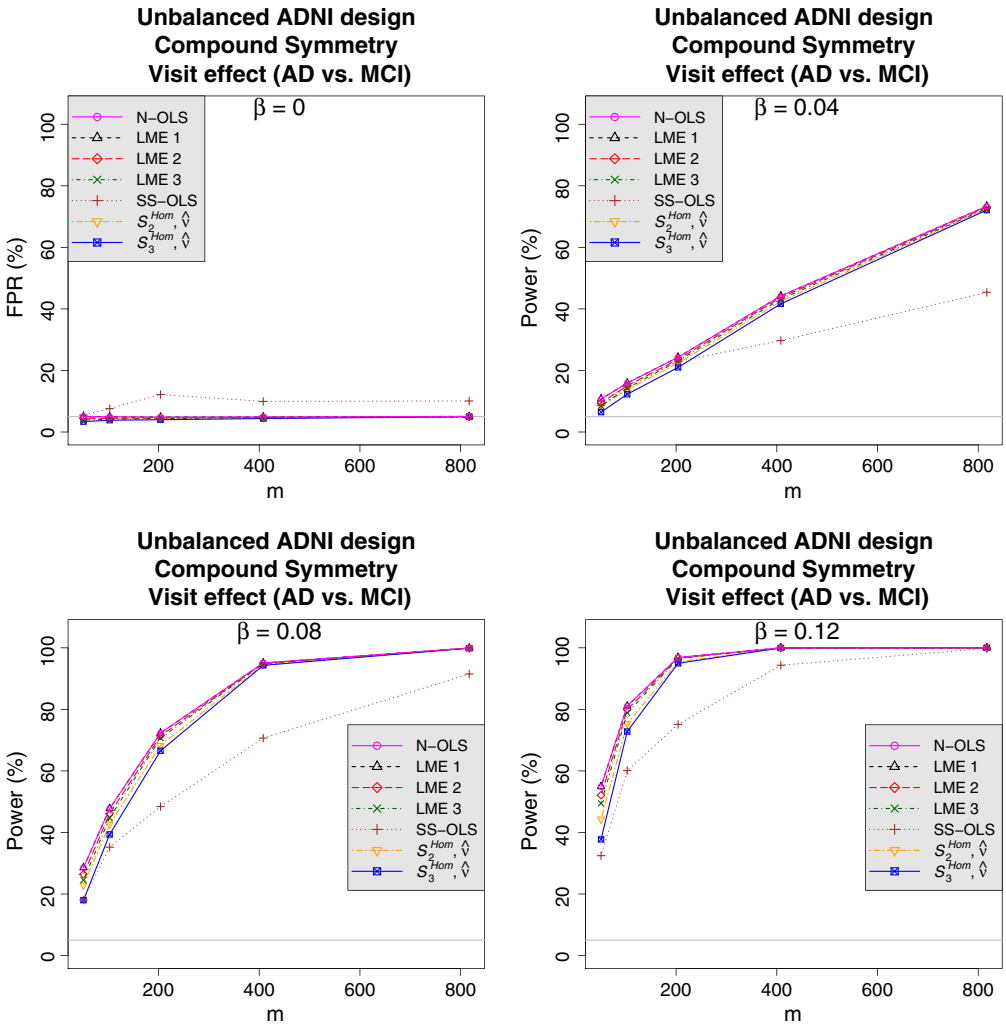

Fig. 3. Power with Compound Symmetry ($\rho = 0.95$) for the unbalanced ADNI design, for varying effect sizes. The tested effect is the difference in the visit effect between AD and MCI groups. All results are based on an F-test at nominal level 5%. LME 1, LME 2 and LME 3 correspond to the LME model including a random intercept per subject, the LME model with a random intercept and a random effect of time per subject and the LME model with a random intercept, a random effect of time and a quadratic effect of time per subject, respectively. See Fig. 1 for a description of the SwE versions.

less powerful than the SwE approach. This effect may seem counterintuitive, but is explained by noting that the SS-OLS method tends to be less efficient than the SwE method (i.e. that the true variance of the parameters obtained with the SS-OLS method tends to be higher than the true variance of the parameters obtained with the SwE). This can be confirmed by computing the Monte Carlo estimates of the true variances of the SS-OLS and SwE methods and comparing them. For the particular setting of Fig. 3 considering the full ADNI design, it appears that the true variance is 2.9 times bigger for the SS-OLS method than the SwE method. As a consequence, the power of the SS-OLS method will increase more slowly than the one of the SwE method when the effect size increases. Thus, provided that the effect size is large enough to overcome the invalid additional power due to the liberal behaviour of the SS-OLS method, the SwE method will be more powerful than the SS-OLS method, as observed in Fig. 3.

While these results highlight the principal weakness of the SwE method, i.e. a reduced power at low m, we stress that these results are only for CS with no variance heterogeneity. When CS or variance homogeneity cannot be safely assumed, only the SwE or LME (with appropriate random effects or covariance structure for the error terms) methods can provide valid inferences.

Note that a power analysis for a greater visit effect in AD relative to MCI subjects under the assumption of a Toeplitz covariance structure can be found in the Web Supplementary Material. For this case, only the SwE (S_3^{Hom}) seemed to be able to control accurately the FPR, with the N-OLS and random-intercept LME methods appearing highly liberal and the SS-OLS and the two richer LME methods appearing slightly liberal, making them invalid. An interesting observation about these results is that the SS-OLS method seemed to be slightly more or equally as powerful as the SwE method, contradicting the observation made in the CS case (see Fig. 3). Comparing the Monte Carlo estimates of the true variances of each method showed that the SS-OLS method is 1.2 times larger than the SwE method, indicating that the SS-OLS method should be less powerful than the SwE method, like in the CS case. Nevertheless, in this setting, the SS-OLS method actually underestimates the variance and in turn inflates the test statistic to such a degree that the SS-OLS is slightly more powerful.

LME convergence failure rates

Regarding the convergence failure experiment (see, subsection Simulations IV), the `lmer` function did not exhibit any convergence

failures. However, the `lme` function exhibited a high rate of convergence problems in many designs. The detailed results about the convergence failures can be found in the Web Supplementary Material.

Real ADNI analysis

Prior to the analysis of the real ADNI data, we conducted a Box's test of Compound Symmetry as described in subsection Box's test of Compound Symmetry with a reduced dataset of 483 subjects who were all scanned at screening and followed up at 6, 12 and 24 months. After controlling for a False Discovery Rate of 5% (using a Bonferroni correction at level 5%, respectively), 97% (56%, respectively) of the voxels survived the thresholding indicating a strong evidence of non-CS in the data.

Fig. 4 compares the t-score images obtained by the N-OLS, SwE (S_3^{Hom}) and SS-OLS methods with the real images for contrasts on the difference between groups in terms of visit effect on the brain atrophy (all methods thresholded at 5 for comparison). The N-OLS method has larger t-values and more supra-threshold voxels than the SwE method. While this could be attributed to power differences, with 817 subjects, we expect negligible differences in power. Hence a more likely explanation is the presence of a complex (non-CS) longitudinal covariance structure that results in inflated significance (Fig. 2, top left and bottom). The SS-OLS has smaller t-values and fewer supra-threshold voxels than the SwE method, likely attributable to conservativeness (Fig. 2, right) and/or reduced power (Fig. 3, top left and bottom).

Figs. 5, 6 and 7 shows the regression fits for three particular voxels situated in different areas of the brain. Note that these voxels were not selected based on maximal difference between the SwE and N-OLS (or SS-OLS) methods, but rather based on relatively high significance in term of age, visit or acceleration effects in all of the methods (qualitatively, the statistic maps for the three methods are similar). As a reminder from subsection Real data analysis, all the scans represent the relative difference in brain volume from the MDT reference image, as such, a value of 10% in the plots indicates that the brain volume is 10% bigger than in the MDT image. Fig. 5 shows results for a voxel in the right anterior cingulate where there is strong evidence of brain atrophy with age and also with the visit effect. The rate of brain atrophy seems similar for each group and is similar for both the age and the visit effect, indicating consistent cross-sectional and longitudinal volume changes. Fig. 6 shows a voxel in the right ventricle where there is strong evidence of an expansion in volume. As expected, this is greater in AD subjects than in MCI or Normal subjects. Fig. 7 shows a voxel in the right posterior cingulate where we observe strong brain atrophy for the AD subjects compared to the Normal subjects. In Figs. 5, 6 and 7, the Normal subjects have similar intra- and inter-subject effects of time (visit and age effects, respectively), and we generally observe this throughout the brain. In contrast, in the AD and MCI groups, there are inconsistent longitudinal and cross-sectional effects of time. Specifically, there is evidence of a "deceleration", where the oldest patients exhibit reduced rates of expansion (or contraction) relative to younger patients. One interpretation is a "saturation" effect, where, with advancing disease progress, there is less gray matter left to atrophy and less space in the cranial vault for the ventricles to expand. However, as the ADNI only follows subjects for at most 3 years, an alternative interpretation must be considered. Specifically, instead of this deceleration reflecting an aspect of the disease process, it rather reflects age-dependent heterogeneity in the ADNI cohort. For example, MCI subjects in their 80's are likely to have systematic differences from the MCI subjects in their 60's, as the

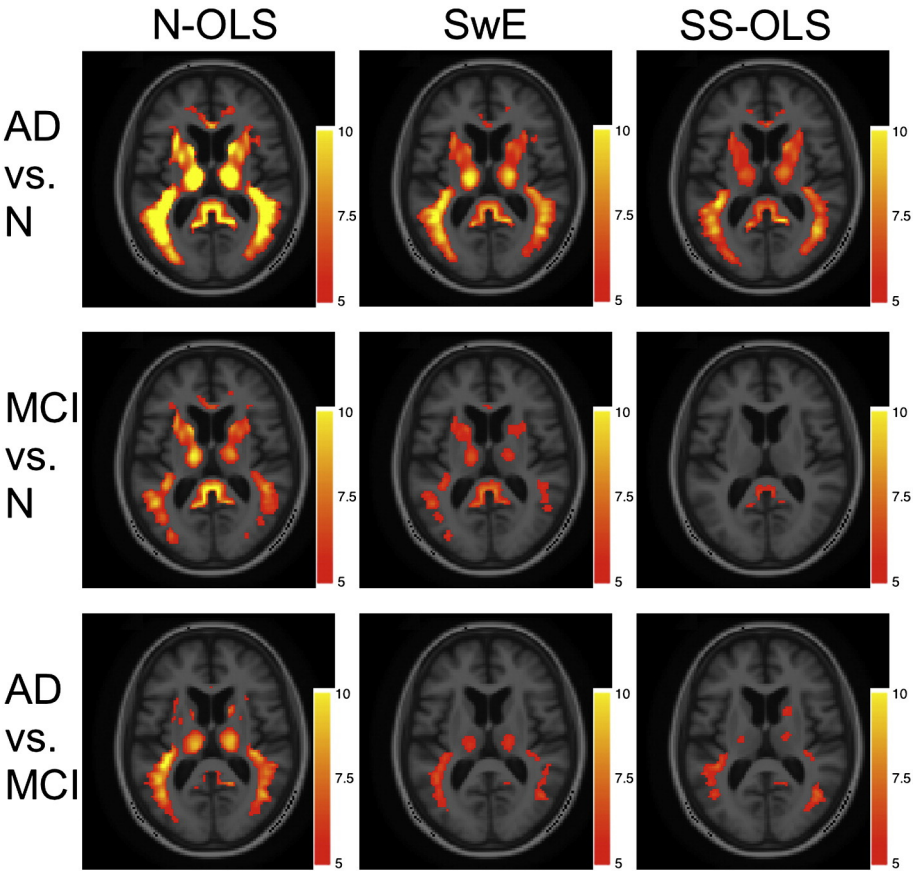

Fig. 4. Thresholded one-sided *t* images for the differential visit effect, greater decline in volume in AD relative to N, MCI relative to N and AD relative to MCI, for the N-OLS, SwE (S_3^{Hom}) and SS-OLS methods; threshold of 5 used for all methods; axial section shown at $z = 14$ mm. Apparent superior sensitivity of the N-OLS method (left) is likely due to inflated significance and poor FPR control; see text and Fig. 2.

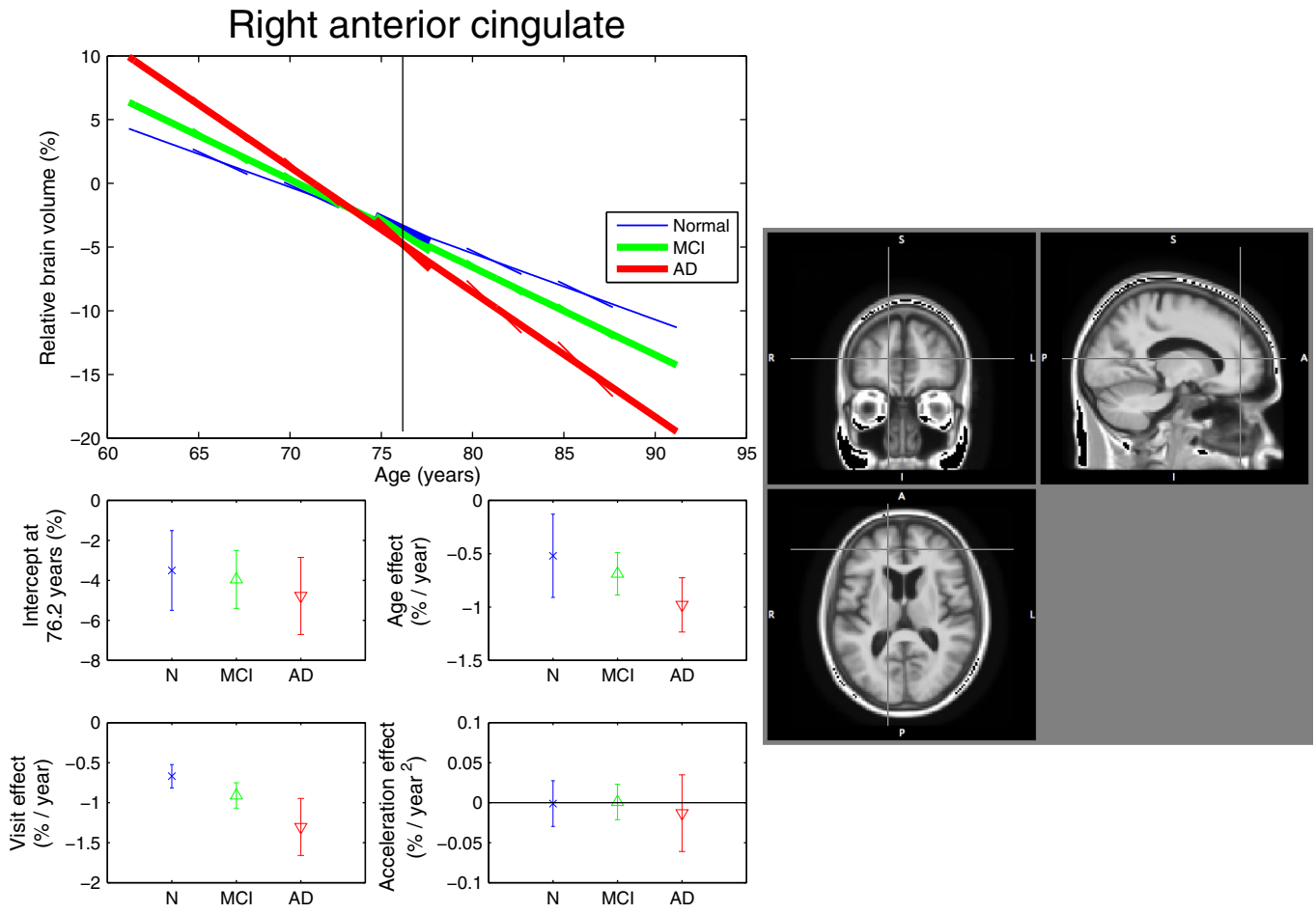

Fig. 5. Model fit in the right anterior cingulate cortex. Top plot: linear regression fit obtained with the SwE method (S_3^{Hom}) at voxel $(x, y, z) = (16, 45, 14)$ mm; the vertical line at 76.2 years marks the average age of the study participants; the thickness of the lines reflects the strength of the t-scores obtained for the age effect (the three main lines), the visit effect (the three secondary lines centred at 76.2 years) and the acceleration effect (the secondary lines centred at 66.2, 71.2, 81.2 and 86.2 years). Bottom plots: 95% confidence intervals for all the parameters of the linear regression. Right image: location of the selected voxel. The confidence intervals suggest that the rate of brain atrophy seems similar for each group and is similar for both the age and the visit effect, indicating consistent cross-sectional and longitudinal volume changes.

former group have survived to their 8th decade free of severe dementia, while some of the latter group will convert to AD in the next 20 years. As pointed out by one of the reviewers, this kind of explanation has already been reported in Thompson et al. (2011).

Computation time

As suggested by one of the reviewers, we compared the elapsed computation times of the SwE and LME methods obtained on a 2.7 GHz quad-core Intel Core i7 MacBook Pro with 16 GB of memory. For this, we considered the scenario where we would like to analyse the 336,331 in-mask voxels of the ADNI dataset (see subsection Real data analysis) with the two methods in R and test for the presence of a visit effect (AD vs. N subjects). Table 5 shows the results obtained with the SwE version S_3^{Hom} (SwE in the table), the LME model including a random intercept per subject (LME 1 in the table), the LME model with a random intercept and a random effect of time per subject (LME 2 in the table), and the LME model with a random intercept, a random effect of time and a quadratic effect of time per subject (LME 3 in the table). Note that our home built R implementation of the SwE method uses four different functions. The first one computes voxel-independent variables which need to be computed only once for the whole brain; the second one computes voxel-specific estimates of β, Var(β) and other variables needed for the estimation of v; the third one computes contrast-specific and voxel-independent variables needed for the estimation of v; and the fourth one computes contrast- and voxel-specific estimates of v. For the LME models, the (voxel-specific) `lmer`, (voxel-specific) `vcovAdj` and (contrast- and voxel-specific) `get_ddf_Lb` functions were used for each voxel.

Discussion

While the SwE is an ubiquitous biostatistical tool, to our knowledge, we are the first authors to provide a detailed study of its small sample properties in a range of settings important for neuroimaging and identify a *non-iterative* estimator that works well for the analysis of longitudinal neuroimaging data.

We have shown that the SwE method is a flexible computationally efficient alternative to the N-OLS, SS-OLS and LME methods. When the simplest covariance structure, CS, cannot be assumed, the SwE (S_3^{Hom}) method and the LME method using appropriate random effects to model correctly the true covariance structure were the only methods that consistently controlled the FPR. In particular, the SS-OLS method was not able to control the FPR in the ADNI design. This effect can be explained by the fact that an inhomogeneity in the distribution of the summary statistics is likely to occur when subjects do not have the same number of observations, leading to a lack of control of the FPR as observed in our simulations. We also have shown that the N-OLS, SS-OLS

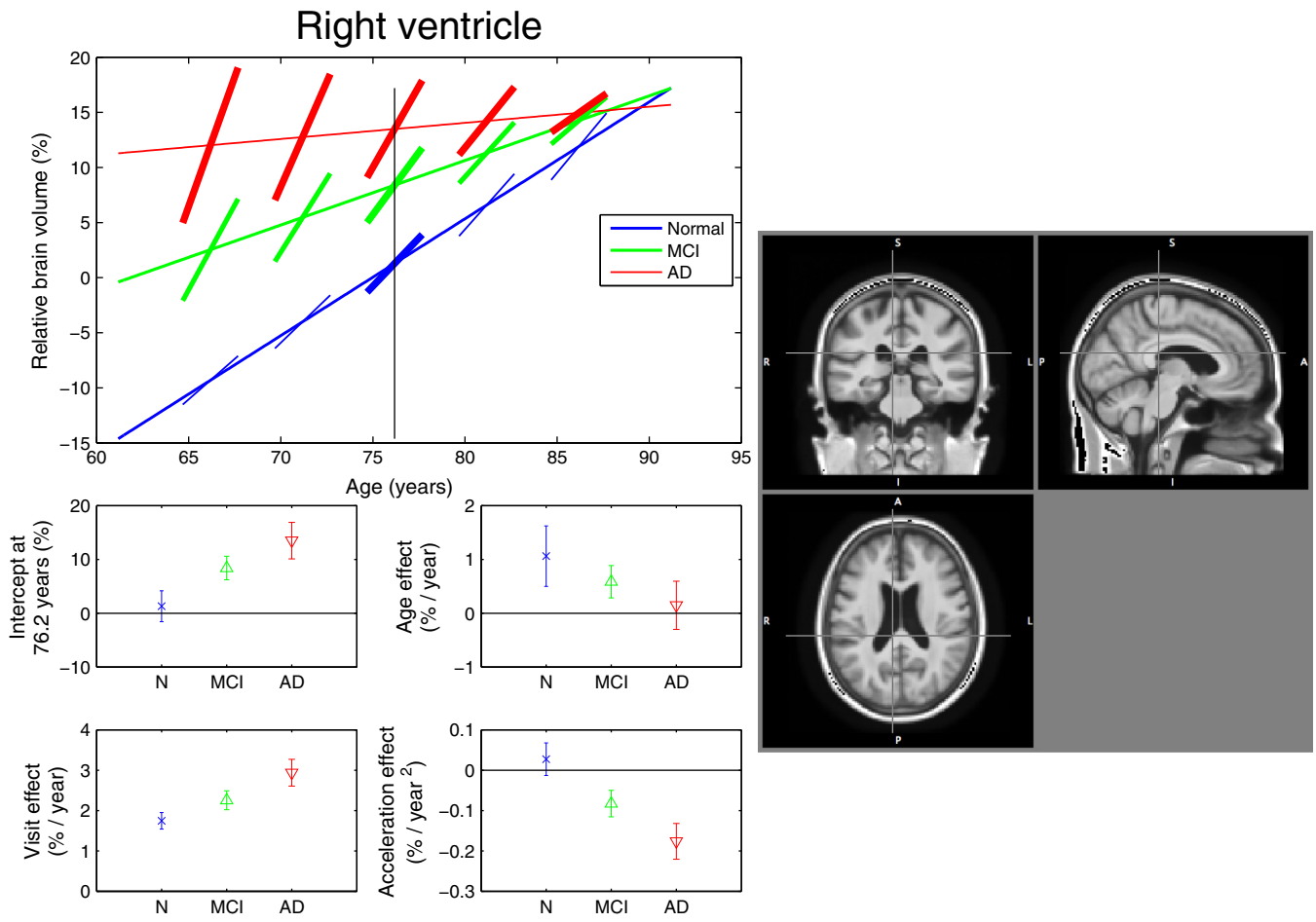

Fig. 6. Model fit in the right ventricle. Top plot: Linear regression fit obtained with the SwE method (S_3^{Hom}) for voxel $(x, y, z) = (8 - 35, 24)$ mm. (See Fig. 5 caption for a description of the different figure components). In the AD and MCI groups a mismatch is observed between cross-sectional and longitudinal effects of time, with a reduced rate of change with increasing age; see body text for more discussion.

and LME methods may be inaccurate when there exists heterogeneity in group variance. Nevertheless, it is worth noting that all of these methods can be adapted to accommodate such a heterogeneity by, for example, specifying different variances for each group in their model. In the SwE method, the use of a marginal model simplifies the specification of the predictors and the interpretation of parameters. In particular, both within- and between-subject covariates can be used, and we have illustrated the ease with which cross-sectional and longitudinal time effects can be used. In particular, testing the interaction of these two time effects revealed a "deceleration" effect in the MCI and AD patient groups that was missing from the healthy controls. We have noted, however, the importance of replacing an arbitrary covariate with two, one purely within-subject and one purely between-subject.

We note that, with our focus on structural data, we did not investigate one-sample t-tests on subject summary statistics. While one-sample t-tests have been shown to be robust under heterogeneity (Mumford and Nichols, 2009), these methods are however less flexible than other regression methods which allow for the inclusion of covariates. Another approach not investigated in this manuscript and which is implemented in SPM12, first estimates a common covariance matrix structure for the whole brain and assumes it to be the true covariance structure for all the voxels in the brain. While there are likely voxels where this common covariance structure is valid, in order to safely use this approach, tests for the accuracy of the assumed covariance should be examined.

In this manuscript, we have also made a comparison between the computation times needed by the SwE method (see subsection Computation time) compared to the LME method, demonstrating the computational efficiency of the SwE method. Nevertheless, it is worth noting that the R implementation of the LME method does not make use of any voxel-independent pre-computations as we used for the SwE method, and thus the LME method could potentially be accelerated. Also, the computation time of the Kenward–Roger covariance matrix correction and the Kenward–Roger effective degrees of freedom were surprisingly high, indicating a likely inefficient implementation in the `pbkrtest` R package. This seems to indicate that the computation time of the LME models could be reduced. Nevertheless, we doubt that this reduction would be large enough to match the computational efficiency of the SwE method.

We have discussed the use of the Box's test for CS, and found ample evidence that the ADNI data's covariance structure is inconsistent with CS.

The principal limitation of the SwE method regards power. When CS holds, it has slightly inferior power to the LME and N-OLS methods, and the recommended S_3^{Hom} SwE was sometimes slightly conservative for samples smaller than 50 in a balanced design and 200 in the highly unbalanced ADNI design. However, when CS doesn't hold, or when there is variance heterogeneity, the N-OLS, SS-OLS and random-intercept LME fail to control False Positives and are *unusable*. Thus, this conservativeness seems like a reasonable price to pay for validity. Also, even when

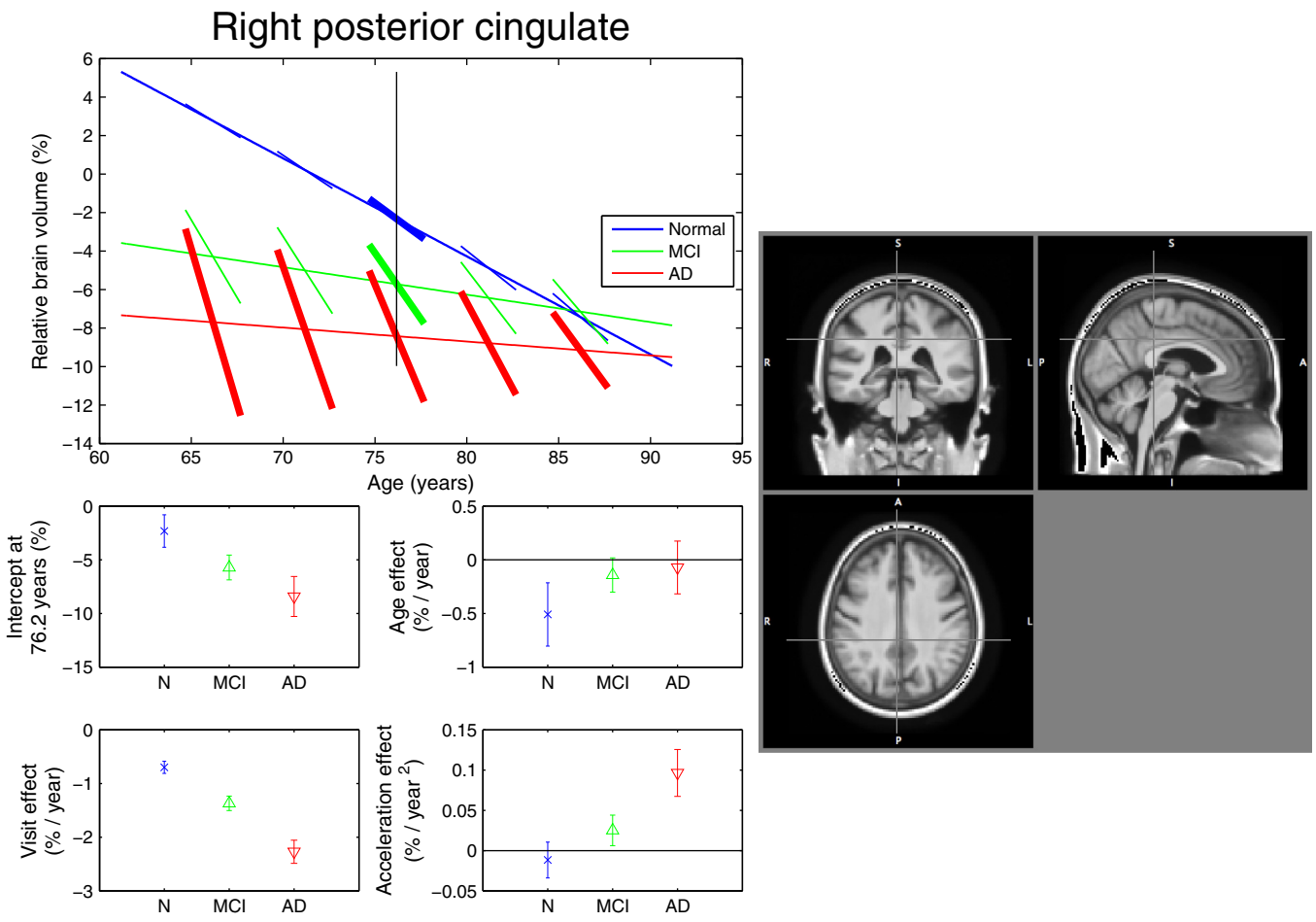

Fig. 7. Model fit in the right posterior cingulate. Top plot: Linear regression fit obtained with the SwE method (S_3^{Hom}) for voxel $(x, y, z) = (4, -39, 38)$ mm. (See Fig. 5 caption for a description of the different figure components). In the AD and MCI groups, there is a mismatch between cross-sectional and longitudinal effects of time, with a reduced rate of change with increasing age; see body text for more discussion.

CS holds, it may be desirable to use the SwE method over the N-OLS method to allow fitting of a mix of within- and between-subject covariates.

If more power is needed, one can use some form of spatial regularisation or more complicated models like in Skup et al. (2012), Bernal-Rusiel et al. (2013b) or Li et al. (2013). Nevertheless, while those methods are expected to be more powerful, they require iterative algorithms, which makes them slower than the SwE method. Moreover, there is no evidence that, at least in some settings, they will do this with a good control of the FPR. Notably, Zhang (2008) showed that using a spatial regularisation will tend to decrease the variance of the estimates (which will tend to increase the power), but also increase their bias (which will tend to alter the accuracy).

It would be desirable to use permutation methods (see, e.g., Nichols and Holmes, 2002) in combination with the SwE to produce non-parametric inferences. However, permutation tests assume that the scans are exchangeable under the null hypothesis, incompatible with longitudinal or repeated measures data. Bootstrap methods (see, e.g., Efron and Tibshirani, 1994), in contrast, do not require the exchangeability assumption and may be applicable. As there are different types of bootstrap tests to consider and extensive small-sample simulations needed to validate this asymptotic method, we have left this for future study.

As another future direction, we intend to check the validity of the Random Field Theory (see, e.g., Worsley et al., 1996) with the SwE method. It is indeed not guaranteed that the assumptions required by the Random Field Theory hold when the SwE method is used. As such, at present, we can only recommend the use of a False Discovery Rate control in order to deal with the multiple comparison problem.

While the present work was motivated and illustrated on a longitudinal dataset, we stress that the SwE can be used to analyse other types of correlated data encountered in neuroimaging. For example, it can be

Table 5
Estimated computation times in days, hours, minutes and seconds in the scenario where the 336,331 in-mask voxels of the TBM ADNI dataset would be tested for an effect of visit (AD vs. N subjects) in R. The setting used corresponded to the one of the second set of simulations (see subsection Simulations II). "n/a", "ind." and "spec." stands for not applicable, independent and specific, respectively; "KR voxel specific" corresponds to the use of the function `vcovAdj`; see text for additional detail.

Computation level	LME 1	LME 2	LME 3	SwE
Voxel-ind.	n/a	n/a	n/a	0d 0 h 0′ 3″
Voxel-spec.	0d 7 h 41′ 57″	1d 1 h 55′ 22″	9d 6 h 5′ 50″	0d 0 h 7′ 11″
KR voxel-spec.	66d 13 h 28′ 44″	111d 20 h 56′ 57″	213d 23 h 46′ 28″	n/a
Contrast-spec. and voxel-ind.	n/a	n/a	n/a	0d 0 h 0′ 1″
Contrast- and voxel-spec.	71d 8 h 57′ 9″	112d 13 h 45′ 56″	215d 21 h 38′ 46″	0d 0 h 0′ 30″
Total	138d 6 h 7′ 50″	225d 12 h 38′ 15″	439d 3 h 31′ 4″	0d 0 h 7′ 44

used to analyse cross-sectional fMRI studies where multiple contrasts of interests are jointly modelled or cross-sectional family studies where subjects from the same family cannot be assumed independent.

Finally, for the real data analysis, the N-OLS, SS-OLS and SwE methods show clearly different results with the SwE method finding fewer significant voxels than the N-OLS method, but more than the SS-OLS method. This seems to be in accordance with our non-CS simulations in which the N-OLS method poorly controls the FPR (and thus has inflated significance; Fig. 2) and the SS-OLS method which is less powerful than the SwE method (Fig. 3). In the simulations, the SwE was accurate for all the different type of covariance structure tested and this seems to make the SwE one of the most trustworthy methods for the analysis of the ADNI data. An SPM extension implementing the SwE method has been made available for use from the authors (http://warwick.ac.uk/tenichols/SwE).

Acknowledgements

B.G. is supported by the Marie Curie Initial Training Network *Methods in Neuroimaging* under grant number MC-ITN-238593. T.E.N. is supported by the Wellcome Trust and NIH grant R01 NS075066-01A. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research& Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organisation is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Appendix A. Approximate test with the SwE

Let $U_1, U_2, \ldots, U_l$ be a sequence of p-dimensional random variables, each independently distributed as a Wishart distribution $W_p(\nu_i, \Sigma_i/\nu_i)$. Nel and Van der Merwe (1986) showed that

$$\sum_{i=1}^{l} U_i \dot{\sim} W_p(\nu, \Sigma/\nu), \tag{A.1}$$

where $\dot{\sim}$ stands for "is approximately distributed as", $\Sigma = \sum_{i=1}^{l} \Sigma_i$ and,

$$\nu = \frac{\mathrm{tr}\left(\Sigma^2\right) + (\mathrm{tr}(\Sigma))^2}{\sum_{i=1}^{l} \frac{\mathrm{tr}\left(\Sigma_i^2\right) + (\mathrm{tr}(\Sigma_i))^2}{\nu_i}}. \tag{A.2}$$

where tr is the trace operator. We will refer to this result as the NVDM approximation.

In the context of a homogeneous version of the SwE with n_G groups, if there is no missing data, V_{0g} is estimated by

$$\hat{V}_{0g} = \frac{1}{m_g} \sum_{i \in \mathcal{I}(g)} e_i^* e_i^{*'} \tag{A.3}$$

where $\mathcal{I}(g)$ is the subset of subjects belonging to group g and e_i^* is an adjusted version of the residuals of subject i. If each e_i^* is correctly adjusted in such a way that each covariance matrix $\mathrm{var}(e_i^*)$ is equal to the covariance matrix of its corresponding true error term $\mathrm{var}(\epsilon_i)$, then they can be assumed to follow a Normal distribution with mean 0 and variance V_{0g} for all $i \in \mathcal{I}(g)$. Then, for all $i \in \mathcal{I}(g)$, we would have

$$B_i = \frac{1}{m_g} e_i^* e_i^{*'} \sim W_{k_g}\left(1, V_{0g}/m_g\right) \tag{A.4}$$

by the definition of a Wishart distribution (Härdle and Simar, 2007), where k_g is the size of e_i^*. If the different subjects' residuals e_i^* were independent, we would have

$$\hat{V}_{0g} = \sum_{i \in \mathcal{I}(g)} B_i \sim W_{k_g}\left(m_g, V_{0g}/m_g\right), \tag{A.5}$$

by the additive property of Wishart distributions. However, this is not the case due to covariates shared between subjects. To account for this dependence, let us first consider a $n \times p$ design matrix X that is separable into n_X sub-design matrices X_u of size $n_u \times p$ such that, defining A_u as the set of non-zero columns in X_u, the collection of sets $\{A_u: u = 1,\ldots, n_X\}$ is pairwise disjoint. Further, let X be composed of p_B pure between-subject covariates (e.g., group intercept, cross-sectional effect of age) and p_W pure within-subject (e.g., longitudinal effect of visit) as recommended in subsection Construction of the design matrix. In such a situation, the residuals e_i^* can be considered to be in a space of dimension $m_i - p_{Bi}$ where m_i is the number of subjects included in the sub-design matrix containing subject i and p_{Bi} is the number of pure between-subject covariates in this sub-design matrix that are not all-zero. Now, we treat the B_i's as independent random variables following a Wishart distribution $W_{k_g}\left(\nu_i, V_{0g}/\left(m_g \nu_i\right)\right)$ with effective degrees of freedom ν_i that are estimated by $1 - p_{Bi}/m_i$. Then, using the NVDM approximation, we get

$$\hat{V}_{0g} \dot{\sim} W_{k_g}\left(\nu_g, V_{0g}/\nu_g\right) \tag{A.6}$$

where

$$\nu_g = \frac{m_g^2}{\sum_{i \in \mathcal{I}(g)} \frac{1}{\nu_i}}. \tag{A.7}$$

Now, let us consider the test $\mathcal{H}_0 : C\beta = 0$ where C is a matrix (or a vector) of rank q defining a combination of parameters (contrast). Contrasting the SwE S with C, we have

$$CSC' = C\left(\sum_{i=1}^{m} X_i' X_i\right)^{-1} \left(\sum_{g=1}^{n_G} \sum_{i \in \mathcal{I}(g)} X_i' \hat{V}_{0g} X_i\right) \left(\sum_{i=1}^{m} X_i' X_i\right)^{-1} C' \tag{A.8}$$

$$= \sum_{g=1}^{n_G} \left(C\left(\sum_{i=1}^{m} X_i' X_i\right)^{-1} \left(\sum_{i \in \mathcal{I}(g)} X_i' \hat{V}_{0g} X_i\right) \left(\sum_{i=1}^{m} X_i' X_i\right)^{-1} C' \right) \tag{A.9}$$

$$= \sum_{g=1}^{n_G} \left(CSC'\right)_g \tag{A.10}$$

where $(CSC')_g$ is the contribution of group g to the contrasted SwE CSC' and which can be rewritten as

$$(CSC')_g = \sum_{i\in\mathcal{I}(g)} D_i \hat{V}_{0g} D_i' \tag{A.11}$$

where

$$D_i = C\left(\sum_{j=1}^{m} X_j' X_j\right)^{-1} X_i'. \tag{A.12}$$

Then, for all $i \in \mathcal{I}(g)$, we get

$$D_i \hat{V}_{0g} D_i' \dot{\sim} W_q\left(\nu_g, D_i V_{0g} D_i'/\nu_g\right) \tag{A.13}$$

where q is the rank of C. As each component $D_i \hat{V}_{0g} D_i'$ is obtained with the same estimate $\hat{V}_{0g}$, there is no contribution of additional degrees of freedom and thus

$$(CSC')_g \dot{\sim} W_q\left(\nu_g, \sum_{i\in\mathcal{I}(g)} D_i V_{0g} D_i'/\nu_g\right). \tag{A.14}$$

Assuming that the contributions $(CSC')_g$'s are independent, using the NVDM approximation and noting that $\sum_{g=1}^{n_G}\sum_{i\in\mathcal{I}(g)} D_i V_{0g} D_i' = C\mathrm{var}\left(\hat{\beta}\right)C'$, we get

$$CSC' = \sum_{g=1}^{n_G} (CSC')_g \dot{\sim} W_q\left(\nu, C\mathrm{var}\left(\hat{\beta}\right)C'/\nu\right). \tag{A.15}$$

where

$$\nu = \frac{\mathrm{tr}\left(\left(C\mathrm{var}\left(\hat{\beta}\right)C'\right)^2\right) + \left(\mathrm{tr}\left(C\mathrm{var}\left(\hat{\beta}\right)C'\right)\right)^2}{\sum_{g=1}^{n_G} \frac{\mathrm{tr}\left(\left(\sum_{i\in\mathcal{I}(g)} D_i V_{0g} D_i'\right)^2\right) + \left(\mathrm{tr}\left(\sum_{i\in\mathcal{I}(g)} D_i V_{0g} D_i'\right)\right)^2}{\nu_g}}. \tag{A.16}$$

Noting that $C\hat{\beta}/\sqrt{\nu} \sim N\left(0, C\mathrm{var}\left(\hat{\beta}\right)C'/\nu\right)$ and assuming that $C\hat{\beta}$ and CSC' are independent, we obtain

$$\nu\left(C\hat{\beta}/\sqrt{\nu}\right)'(CSC')^{-1}\left(C\hat{\beta}/\sqrt{\nu}\right) \dot{\sim} \frac{\nu q}{\nu - q + 1} F(q, \nu - q + 1) \tag{A.17}$$

and we finally get the test statistic

$$\frac{\nu - q + 1}{\nu q}\left(C\hat{\beta}\right)'(CSC')^{-1}\left(C\hat{\beta}\right) \dot{\sim} F(q, \nu - q + 1). \tag{A.18}$$

The extension to the heterogeneous SwE case is straightforward as it is equivalent to the homogeneous SwE considering m groups composed by a single subject. In practice, $\mathrm{var}\left(\hat{\beta}\right)$ and V_{0g}'s are unknown, thus, their estimates S and $\hat{V}_{0g}$'s are used instead in Eq. (A.16) to get an estimation of ν. When a group has a very small number of subjects, it will produce poor variance estimates and, consequently, affects the quality of the estimation of ν. This motivates our idea of assuming homogeneous variances between subjects in the computation of the SwE.

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2014.03.029.

References

Bates, D., Maechler, M., Bolker, B., 2012. lme4: Linear mixed-effects models using S4 classes. R package version 0.999999–0. URL http://CRAN.R-project.org/package=lme4.

Bell, R.M., McCaffrey, D.F., 2002. Bias reduction in standard errors for linear regression with multi-stage samples. Surv. Methodol. 28 (2), 169–182.

Bernal-Rusiel, J.L., Greve, D.N., Reuter, M., Fischl, B., Sabuncu, M.R., 2013a. Statistical analysis of longitudinal neuroimage data with linear mixed effects models. NeuroImage 66, 249–260.

Bernal-Rusiel, J.L., Reuter, M., Greve, D.N., Fischl, B., Sabuncu, M.R., 2013b. Spatiotemporal linear mixed effects modeling for the mass-univariate analysis of longitudinal neuroimage data. NeuroImage 81, 358–370.

Box, G.E., 1950. Problems in the analysis of growth and wear curves. Biometrics 6 (4), 362–389.

Chen, G., Saad, Z.S., Britton, J.C., Pine, D.S., Cox, R.W., 2013. Linear mixed-effects modeling approach to fMRI group analysis. NeuroImage 73, 176–190.

Chesher, A., Jewitt, I., 1987. The bias of a heteroskedasticity consistent covariance matrix estimator. Econometrica 1217–1222.

Diggle, P., Liang, K., Zeger, S., 1994. Analysis of longitudinal data oxford statistical science series, 13.

Efron, B., Tibshirani, R.J., 1994. An introduction to the bootstrap (chapman & hall/crc monographs on statistics & applied probability).

Eicker, F., 1963. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. Ann. Math. Stat. 34 (2), 447–456.

Eicker, F., 1967. Limit theorems for regressions with unequal and dependent errors. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 59–82.

Fay, M., Graubard, B., 2001. Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics 57 (4), 1198–1206.

Fitzmaurice, G.M., 1995. A caveat concerning independence estimating equations with multivariate binary data. Biometrics 309–317.

Halekoh, U., Højsgaard, S., 2013. pbkrtest: Parametric bootstrap and Kenward Roger based methods for mixed model comparison. R package version 0.3–8. URL http://CRAN.R-project.org/package=pbkrtest.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 1029–1054.

Hardin, J., 2001. Small sample adjustments to the sandwich estimate of variance. http://www.stata.com/support/faqs/stat/sandwich.html.

Härdle, W., Simar, L., 2007. Applied Multivariate Statistical Analysis. Springer Verlag.

Harville, D., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. 320–338.

Hinkley, D., 1977. Jackknifing in unbalanced situations. Technometrics 285–292.

Horn, S., Horn, R., Duncan, D., 1975. Estimating heteroscedastic variances in linear models. J. Am. Stat. Assoc. 380–385.

Hua, X., Hibar, D.P., Ching, C.R., Boyle, C.P., Rajagopalan, P., Gutman, B.A., Leow, A.D., Toga, A.W., C. R. J. Jr., Harvey, D., Weiner, M.W., Thompson, P.M., 2013. Unbiased tensor-based morphometry: improved robustness and sample size estimates for Alzheimer's disease clinical trials. NeuroImage 66 (0), 648–661.

Huber, P., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, vol. 1, pp. 221–233.

Kauermann, G., Carroll, R., 2001. A note on the efficiency of sandwich covariance matrix estimation. J. Am. Stat. Assoc. 96 (456), 1387–1396.

Kenward, M.G., Roger, J.H., 1997. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 983–997.

Lai, T.L., Small, D., 2007. Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. J. R. Stat. Soc. Ser. B (Stat Methodol.) 69 (1), 79–99.

Laird, N., Ware, J., 1982. Random-effects models for longitudinal data. Biometrics 963–974.

Li, Y., Gilmore, J.H., Shen, D., Styner, M., Lin, W., Zhu, H., 2013. Multiscale adaptive generalized estimating equations for longitudinal neuroimaging data. NeuroImage 72, 91–105.

Liang, K., Zeger, S., 1986. Longitudinal data analysis using generalized linear models. Biometrika 73 (1), 13–22.

Lindquist, M., Spicer, J., Asllani, I., Wager, T., 2012. Estimating and testing variance components in a multi-level glm. NeuroImage 59 (1), 490–501.

Lipsitz, S., Ibrahim, J., Parzen, M., 1999. A degrees-of-freedom approximation for a t-statistic with heterogeneous variance. J. R. Stat. Soc. Ser. D Stat. 48 (4), 495–506.

Long, J., Ervin, L., 2000. Using heteroscedasticity consistent standard errors in the linear regression model. Am. Stat. 217–224.

MacKinnon, J., White, H., 1985. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. J. Econ. 29 (3), 305–325.

Mancl, L., DeRouen, T., 2001. A covariance estimator for gee with improved small-sample properties. Biometrics 57 (1), 126–134.

McDonald, B.W., 1993. Estimating logistic regression parameters for bivariate binary data. J. R. Stat. Soc. Ser. B Methodol. 391–397.

Molenberghs, G., Verbeke, G., 2011. A note on a hierarchical interpretation for negative variance components. Stat. Model. 11 (5), 389–408.

Mueller, S., Weiner, M., Thal, L., Petersen, R., Jack, C., Jagust, W., Trojanowski, J., Toga, A., Beckett, L., 2005. The Alzheimer's Disease Neuroimaging Initiative. Neuroimaging Clin. N. Am. 15 (4), 869.

Mumford, J.A., Nichols, T., 2009. Simple group fMRI modeling and inference. NeuroImage 47 (4), 1469–1475.

Nel, D., Van der Merwe, C., 1986. A solution to the multivariate Behrens–Fisher problem. Commun. Stat. Theory Meth. 15 (12), 3719–3735.

Neuhaus, J., Kalbfleisch, J., 1998. Between-and within-cluster covariate effects in the analysis of clustered data. Biometrics 638–645.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15 (1), 1–25.
Pan, W., 2001. On the robust variance estimator in generalised estimating equations. Biometrika 88 (3), 901–906.
Pan, W., Wall, M., 2002. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. Stat. Med. 21 (10), 1429–1441.
Pepe, M.S., Anderson, G.L., 1994. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Commun. Stat. Simul. Comput. 23 (4), 939–951.
Pinheiro, J., Bates, D., 2000. Mixed-effects Models in s and s-plus. Statistics and Computing. Springer-Verlag, Berlin, D.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team, 2013. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-113.
R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (URL http://www.R-project.org/).
Skup, M., Zhu, H., Zhang, H., 2012. Multiscale adaptive marginal analysis of longitudinal neuroimaging data with time-varying covariates. Biometrics 68 (4), 1083–1092.
Thompson, W.K., Hallmayer, J., O'Hara, R., 2011. Design considerations for characterizing psychiatric trajectories across the lifespan: application to effects of apoe-e4 on cerebral cortical thickness in Alzheimer's disease. Am. J. Psychiatr. 168 (9), 894–903.
Verbeke, G., Molenberghs, G., 2009. Linear Mixed Models for Longitudinal Data. Springer.
Waldorp, L., 2009. Robust and unbiased variance of glm coefficients for misspecified autocorrelation and hemodynamic response models in fmri. J. Biomed. Imaging 15.
West, B., Welch, K.B., Galecki, A.T., 2006. Linear Mixed Models: A Practical Guide Using Statistical Software. CRC Press.
White, H., 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 817–838.
Worsley, K.J., Marrett, S., Neelin, P., Vandal, A.C., Friston, K.J., Evans, A.C., et al., 1996. A unified statistical approach for determining significant signals in images of cerebral activation. Hum. Brain Mapp. 4 (1), 58–73.
Zhang, H., 2008. Advances in Modeling and Inference of Neuroimaging Data. (Ph.D. thesis) The University of Michigan.
Zhao, L.P., Prentice, R.L., Self, S.G., 1992. Multivariate mean parameter estimation by using a partly exponential model. J. R. Stat. Soc. Ser. B Methodol. 805–811.